

Contrasting Patterns of Sequence Evolution at the Functionally Redundant *bric à brac* Paralogs in *Drosophila melanogaster*

Ryan D. Bickel · Wendy S. Schackwitz ·
Len A. Pennacchio · Sergey V. Nuzhdin ·
Artyom Kopp

Received: 9 December 2008 / Accepted: 9 July 2009 / Published online: 29 July 2009
© The Author(s) 2009. This article is published with open access at Springerlink.com

Abstract Genes with overlapping expression and function may gradually diverge despite retaining some common functions. To test whether such genes show distinct patterns of molecular evolution within species, we examined sequence variation at the *bric à brac* (*bab*) locus of *Drosophila melanogaster*. This locus is composed of two anciently duplicated paralogs, *bab1* and *bab2*, which are involved in patterning the adult abdomen, legs, and ovaries. We have sequenced the 148 kb genomic region spanning the *bab1* and *bab2* genes from 94 inbred lines of *D. melanogaster* sampled from a single location. Two non-coding regions, one in each paralog, appear to be under selection. The strongest evidence of directional selection is found in a region of *bab2* that has no known functional role. The other region is located in the *bab1* paralog and is known to contain a *cis*-regulatory element that controls sex-specific abdominal pigmentation. The coding region of *bab1* appears to be under stronger functional constraint than the *bab2* coding sequences. Thus, the two paralogs are evolving under different selective regimes in the same natural population, illuminating the different evolutionary trajectories of partially redundant duplicate genes.

Keywords *Drosophila melanogaster* · *bric à brac* · Population genetics · Pigmentation · Duplicated genes

Introduction

Gene duplication makes an important contribution to the evolution of novel functions and the modifications of existing functions (reviewed in Prince and Pickett 2002), and duplicated genes are prevalent throughout metazoans (Holland et al. 1994; Amores et al. 1998; Force et al. 1999; Holland 1999; Cresko et al. 2003; Amores et al. 2004). Two major theories have been advanced for the maintenance of gene duplications. One theory (the neo-functionalization model) postulates that one of the duplicated genes evolves a novel function while losing some aspect of the ancestral function. Thus, both genes are maintained by natural selection, one for the ancestral function and the other for the new function (Ohno 1970). Alternatively, each gene could accumulate complementary degenerative mutations in either coding or regulatory regions, resulting in the loss of a subset of the pre-duplication gene activity. Both copies will then be maintained by selection since both are needed to preserve the ancestral function (the sub-functionalization model) (Force et al. 1999; Lynch and Force 2000).

Both models predict that, once duplicated genes occupy different functional niches, they may come under different selective regimes. Paralogous gene regions responsible for redundant functions may experience similar selective pressures, if their overall activity affects fitness traits. On the other hand, selection acting on different functions may shape sequence evolution at functionally differentiated regions, and the mode and intensity of this selection may be different for each trait. As a result, duplicated genes

R. D. Bickel (✉) · S. V. Nuzhdin
Section of Molecular and Computational Biology, Department of
Biological Sciences, University of Southern California,
1050 Childs Way, Los Angeles, CA 90089, USA
e-mail: rbickel@usc.edu

A. Kopp
Department of Evolution and Ecology, University of
California—Davis, Davis, CA 95616, USA

W. S. Schackwitz · L. A. Pennacchio
U.S. Department of Energy Joint Genome Institute,
Walnut Creek, CA 94598, USA

would evolve in different modes and at different rates, with different functional elements dominating the evolution of each paralog. These ideas have primarily been tested using recently duplicated genes, but even old duplicates can share some functions. Here, we analyze the *bric à brac* (*bab*) locus of *Drosophila melanogaster*, which contains the duplicated paralogs *bab1* and *bab2* to determine whether we can detect these patterns using intraspecific variation.

bab1 and *bab2* are located in a 148 kb continuous region of the genome, and their transcripts span ~57 and ~28 kb, respectively (Fig. 1). The large size of the *bab* locus and the low levels of linkage disequilibrium in the region (Fig. 2) suggest that separate regulatory modules and coding regions could evolve independently. Both genes function in the proximo-distal patterning of legs and antennae, the development of terminal filaments in the ovary, and the patterning of abdominal sensory organs and abdominal pigmentation (Couderc et al. 2002). Although *bab1* and *bab2* have overlapping and partially redundant roles in development (Kopp et al. 2000; Couderc et al. 2002), the maintenance of these duplicates in all of the currently sequenced *Drosophila* genomes (unpublished analysis) suggests that they may have subtle differences in function.

Many of the structures patterned by the *bab* genes are sexually dimorphic, including the gonad (Sahut-Barnola et al. 1995), the sex combs on the front legs of males (Godt et al. 1993; Barmina and Kopp 2007; Randsholt and Santamaria 2008), ventral abdominal bristles (Kopp et al. 2000), and the dorsal abdominal pigmentation pattern (Kopp et al. 2000; Williams et al. 2008), which suggests that sexual selection may have acted on the *bab* locus. Furthermore, it has been demonstrated that sex combs are important for male mating success (Ng and Kopp 2008), ovaries are critical for reproduction and fecundity, and abdominal pigmentation plays a role in thermoregulation and desiccation resistance (Gibert et al. 1996; Brisson et al. 2005), suggesting that the *bab1* and *bab2* genes may experience selection on a variety of functions.

The *bab1* and *bab2* genes have sequence conservation in two protein domains, BTB/POZ and BabCD (*bric à brac* conserved domain) (Couderc et al. 2002). The BTB (Broad, Tramtrac, Bab) domain is shared by a large number of developmentally regulated genes and is involved in protein–protein interactions (Zollman et al. 1994), including *bab1* homodimerization in vitro (Chen et al. 1995). The BabCD is composed of a Psq and AT-hook domains that are both involved in DNA binding, suggesting that the *bab* genes may act as transcriptional regulators (Reeves and

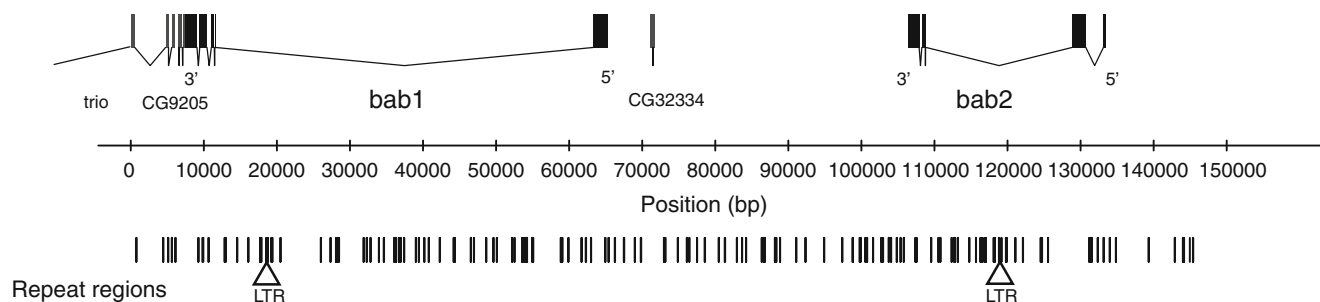
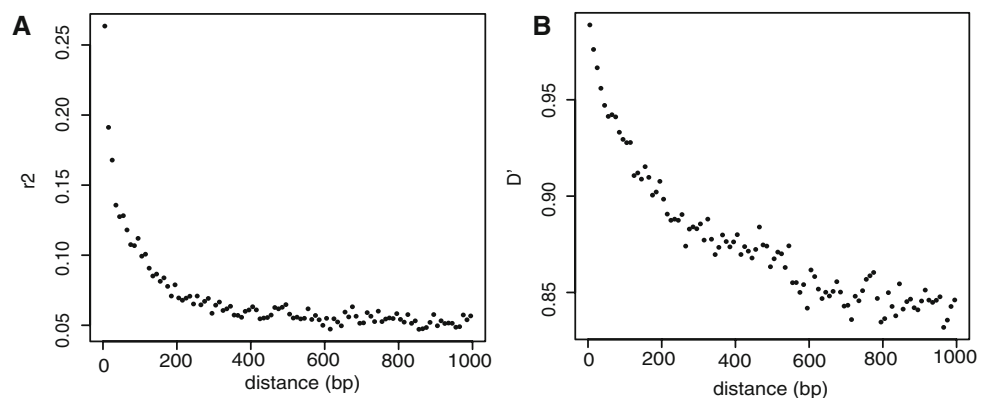


Fig. 1 The *bab* genomic region. The *bab* genes are shown in black, CG32334, CG9205, and the 5' region of *trio* in gray. Repeat regions identified using RepeatMasker are marked below the ruler. The sites

of two transposable element insertions in the *D. melanogaster* reference genome are marked with triangles

Fig. 2 Linkage disequilibrium in the *bab* region. For each graph, the mean value of r^2 (a) and D' (b) was calculated for all polymorphisms separated by a given distance and combined into 10 bp bins



Nissen 1990; Lehmann et al. 1998; Couderc et al. 2002; Lours et al. 2003). Both *bab* genes contain a single large intron (20 and 50 kb, respectively) that is present in an evolutionarily conserved position (Couderc et al. 2002). In both genes, this intron separates 5' exons, which contain the protein interaction domain (BTB) from 3' exons that contain the DNA binding region (BabCD) (Fig. 1).

bab1 and *bab2* have largely overlapping expression patterns, with *bab1* present in a subset of *bab2*-expressing cells. In the ovary, *bab1* is expressed exclusively in the terminal filaments, while *bab2* is expressed strongly in the terminal filaments and more weakly in apical cells of the ovary (Couderc et al. 2002). Flies with *bab* mutations that affect both paralogs show defects in terminal filament formation, apical cells, and basal stalk primordium, resulting in sterile females and ovaries with only a few rudimentary ovarioles, while mutations that affect a single *bab* gene result in weaker phenotypes (Godt and Laski 1995; Couderc et al. 2002). Both duplicate genes also contribute to the patterning of distal antennae and legs during larval and pupal development (Godt et al. 1993; Chu et al. 2002; Couderc et al. 2002). Again, the strongest phenotypes result from *bab* mutations that affect both *bab1* and *bab2*, causing a complete fusion of the second through fifth tarsal segments, while mutations that affect only one of the genes result in intermediate phenotypes.

In the abdomen, the *bab* genes play a central role in specifying sexually dimorphic pigmentation patterns (Kopp et al. 2000). *bab* mutations have a dominant effect resulting in wider pigmentation bands, with the strongest phenotype seen in the most posterior segments (Couderc et al. 2002). Moreover, genetic variation at the *bab* locus is associated with intraspecific variation in the pigmentation of posterior abdominal segments in *D. melanogaster* females (Kopp et al. 2003). *bab1* and *bab2* are expressed in similar spatial patterns in the developing abdominal epidermis (Kopp et al. 2000; Williams et al. 2008), and artificial over-expression experiments show that both genes are capable of partially rescuing the *bab* mutant phenotypes (Bardot et al. 2002). In all tissues, despite slight differences in *bab* expression, *bab1* and *bab2* mutations have very similar phenotypes.

Detailed functional analysis of the *bab* locus has revealed a number of distinct *cis*-regulatory elements (CREs) (Williams et al. 2008). Separate enhancers were identified for pupal abdominal epidermis (large intron of *bab1*), legs (intergenic region between *bab1* and *bab2*), and oenocytes (large intron of *bab2*). Surprisingly, only a single regulatory element was identified for each tissue that expresses both *bab1* and *bab2*, raising the possibility that both paralogs may be controlled by the same “core” CREs. This does not rule out the existence of other, paralog-specific regulatory elements that modulate the expression

of each gene in a more subtle way. If such modifier elements exist, the expression of each paralog could evolve independently and be subject to different selective regimes.

In summary, the two *bab* genes have largely overlapping expression and developmental roles, yet they show evidence of distinct functional specificities. At the same time, their involvement in a variety of sex-specific processes suggests that these genes could experience many competing selective pressures. In principle, both paralogs could be dominated by similar selective pressures, reflecting their shared functions. Alternatively, *bab1* and *bab2* could show different patterns of selection, suggesting that unique functions of the paralogs are shaping sequence evolution in the region. To distinguish between these modes of evolution, we analyzed intraspecific variation throughout the *bab* genomic region.

Materials and Methods

We have resequenced the *bab* genomic region including the *bab1* and *bab2* genes and the flanking intergenic regions from 94 inbred strains extracted from a single natural population at the Wolfskill orchard in Winters, CA. The 35 Wolfskill-1 (W1), 56 Wolfskill-3 (W3), and 3 A1 lines were all collected from the same orchard but in separate years. Eighty-three of the Wolfskill lines were chosen at random, while the remaining lines were chosen for inclusion because of their light abdominal pigmentation pattern. The removal of these lines from the analysis did not significantly change the results. All lines from the Wolfskill collections were inbred by full-sib mating for a minimum of 20 generations, while the A1 lines were inbred for at least 10 generations by the same method.

Sanger based sequencing (ABI 3730xl) was performed at the Joint Genome Institute. Overlapping 1 kb amplicons were designed across the region; successful amplicons were sequenced from both strands. Base calls and polymorphisms were initially identified using Phred and PolyPhred 6.11 (Ewing and Green 1998; Ewing et al. 1998; Stephens et al. 2006). Using Consed, insertion/deletions (indels) were identified and polymorphisms were checked for accuracy (Gordon et al. 1998). Although effort was made to obtain complete coverage, we were unable to sequence any of the strains for two regions that together cover approximately 5 kb. These regions are identified as repetitive by RepeatMasker (Smit 1996–2004), and each region contains a transposable element in the *D. melanogaster* reference genome sequence (Adams et al. 2000). Since transposable elements present in the reference annotation are rarely found in other strains at appreciable frequencies (Petrov et al. in preparation), we did not attempt to verify their presence in our lines. On average,

we have sequence information from 90% of the lines for any given polymorphism.

Sliding window analysis was used to calculate population-genetic test statistics in 10 kb windows that were moved by 2 kb steps across the length of the *bab* region. Theta values (π , θ_W , and θ_H), Tajima's D , Fu and Li's D and Fu's F were calculated using the compute implementation of libsequence library (Thornton 2003) and custom scripts, using the *D. simulans* genome sequence as an outgroup when appropriate (Tajima 1989; Fu and Li 1993; Fay and Wu 2000; Thornton 2003; Zeng et al. 2006). F_{st} was calculated as described in Hudson et al. (1992). Polarized and unpolarized McDonald–Kreitman (MK) tests (McDonald and Kreitman 1991) were performed as described by Begun et al. (2007). All figures were produced using the *R* statistical package (<http://www.R-project.org>).

Results

We sequenced the 148 kb *bab* region from 94 inbred strains of *D. melanogaster* collected from Winters, CA. This region includes *bab1* and *bab2* in their entirety as well as two additional open reading frames, *CG9205* and *CG32334*, for which no information about expression or function is available (Fig. 1). In this region, we identified 5566 single nucleotide polymorphisms (SNPs), 5405 of which contained two alleles and 161 contained three alleles (Table 1). We also identified 1211 short insertion/deletion (indel) polymorphisms, ranging in size from 1 to 526 bp. This is likely to be an underestimate of the number of indels because we have no information from the repeat

regions (Fig. 1) and longer repeat variants from other regions were likely to result in a failed sequencing reaction. The mean indel length was 9.37 bp, and the most common length 4 bp. Although indels are less frequent than SNPs, we find that indels tend to have a higher Tajima's D statistic than SNPs, indicating that they are more likely to be present at intermediate frequencies (Table 1). This result is unlikely to be caused by sequencing errors, since the removal of singletons (SNPs present in a single line) results in the same pattern (data not shown).

Nucleotide diversity (π) and estimates of the population mutation rate (θ_W) were generally higher in non-coding than coding DNA, suggesting that non-coding sequences are under less functional constraint (Table 1). Intronic, intergenic, and UTR regions have similar values of Tajima's D , which is consistent with genome-wide studies in *D. melanogaster* (Andolfatto 2005). The *bab* region shows little linkage disequilibrium (LD), with average correlation between polymorphisms (r^2) dropping off rapidly within 300 bps (Fig. 2a). D' , a quantitative measure of LD normalized for allele frequency (Lewontin 1964), has a slower decline and is constant after 1 kb (Fig. 2b). The short range of LD suggests that different regions of the *bab* locus can, in principle, evolve independently of one another.

To investigate and compare the evolutionary forces acting on the *bab* paralogs, we used several tests to examine the allele frequency spectrum across the *bab* locus. Tajima's D statistic compares two estimators of the population mutation parameter θ : π , a measure of average pairwise differences between sequences that is strongly influenced by common alleles, and θ_W , which weighs all polymorphisms equally and is thus more strongly influenced by rare alleles (Watterson 1975; Tajima 1989). Sliding window analysis shows variable values of Tajima's D statistic across the *bab* locus. A region centered on the non-coding 3' UTR of *bab2* transcript (near the 110,000 bp mark) has negative D values (Fig. 3b), indicating an excess of low frequency alleles that may be due to recent selection. The remainder of the *bab* locus has Tajima's D values near zero, indicating that this region is evolving neutrally.

We also compared high-frequency derived and intermediate-frequency alleles using Fay and Wu's H statistic (Fay and Wu 2000; Zeng et al. 2006). A low value of H indicates a higher than expected number of derived alleles, making it a powerful test for detecting positive selection and the initial stages of balancing selection (Zeng et al. 2006). We used *D. simulans* genome sequence as an outgroup to polarize SNP alleles in *D. melanogaster*. Similar to Tajima's D , the strongest negative values of Fay and Wu's H are found near the 3' end of the *bab2* transcript, with no comparable signature in the paralogous *bab1* region (Fig. 3c). This pattern provides additional evidence for directional selection acting on the region near the 3' end

Table 1 Summary statistics of sequence variation in the *bab* region

	bp ^a	S ^b	Singletons ^c	θ_W^d	π^e	Tajima's D
Entire region	147959	6655	1382	0.90	0.87	−0.13
SNPs		5535	1189	0.75	0.71	−0.18
Indels		1120	193	0.15	0.16	0.15
Coding	8692	292	37	0.67	0.71	0.23
UTRs	4572	156	29	0.68	0.64	−0.22
5'-UTRs	857	26	5	0.61	0.58	−0.17
3'-UTRs	3715	130	24	0.70	0.65	−0.23
Introns	75870	3366	746	0.89	0.84	−0.19
Intergenic	58825	2859	573	0.97	0.95	−0.07

^a Number of base pairs represented in the sample

^b Number of polymorphisms identified

^c The number of polymorphisms that contain a single observation of the alternative allele

^d Watterson's Theta per 100 bp

^e The average pairwise diversity per 100 bp

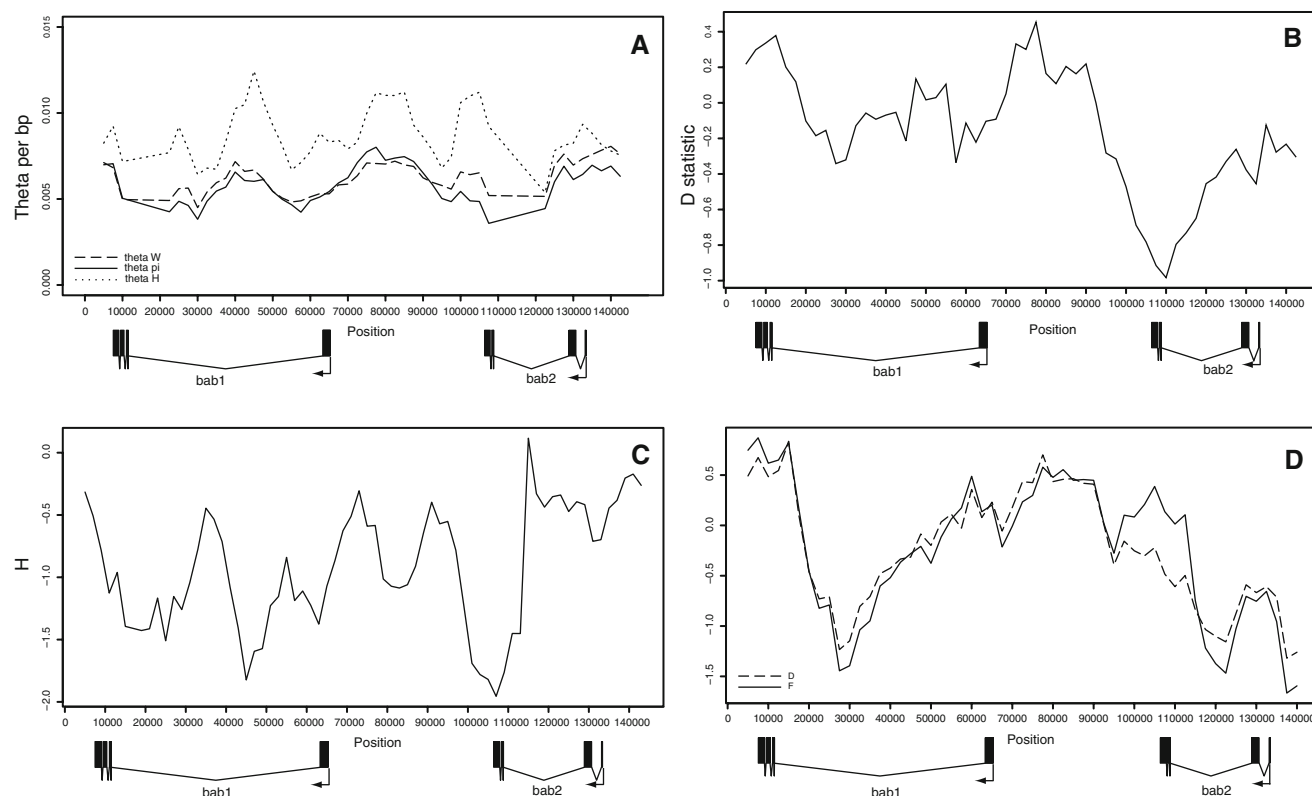


Fig. 3 Sliding window analysis of sequence variation in the *bab* region. All analyses are in 10 kb windows offset by 2 kb. **a** Three estimates of θ including π (solid line), θ_W (dashed line), and θ_H

(dotted line). **b** Tajima's D . **c** Fay and Wu's H statistic. **d** Fu and Li's D (dashed line) and F (solid line) statistics

of *bab2*. In addition, the H statistic shows a region in the large intron of *bab1* with strongly negative values, suggesting an additional region under selection, which was not detected with the D statistic.

Fu and Li's D and Fu's F statistics compare the frequencies of derived and ancestral alleles to detect deviations from the neutral expectation (Fu and Li 1993; Fu 1997). Negative values of D and F indicate an excess of derived mutations (an excess of external branches in the gene tree), while positive values show a deficiency of derived alleles (excess of internal branches). Fu and Li's D is particularly sensitive to background selection—a reduction of diversity at a neutral locus due to selection against linked deleterious mutations (Charlesworth et al. 1993). We find negative values of D and F in the large introns of both *bab1* and *bab2*, with peak values in the 10 kb windows centered near the 28,000 and 122,000 bp marks (Fig. 3d). These regions overlap with the locations of repetitive sequences and transposable element insertions in the reference genome sequence (Fig. 1). The same pattern remains if we repeat the analysis with these repetitive regions masked. Repetitive sequences are often a source of frequently occurring deleterious mutations, and the low values of D and F may arise when these mutations are removed by background selection.

Our sequencing sample is drawn primarily from two collections, Wolfskill 1 (W1) and Wolfskill 3 (W3), which were collected from the same location but in separate years. In general, sequence variation in the W1 and W3 samples does not differ significantly across the *bab* region (Table 2). However, the strong negative values of Tajima's D near the 3' end of *bab2* are caused primarily by the W3 sample (Fig. 4a), while the W1 collection has D values closer to zero. Sliding window analysis of population differentiation (F_{st}) between W1 and W3 reveals elevated levels of differentiation in the same region (Fig. 4b). These

Table 2 Summary statistics comparing the W1 and W3 sequence samples across the *bab* region

	bp ^a	S ^b	Singletons ^c	θ_W^d	π^e	Tajima's D
W1	147959	5323	1347	0.90	0.89	−0.04
W3	147959	5845	1384	0.88	0.87	−0.06

^a Number of base pairs represented in the sample

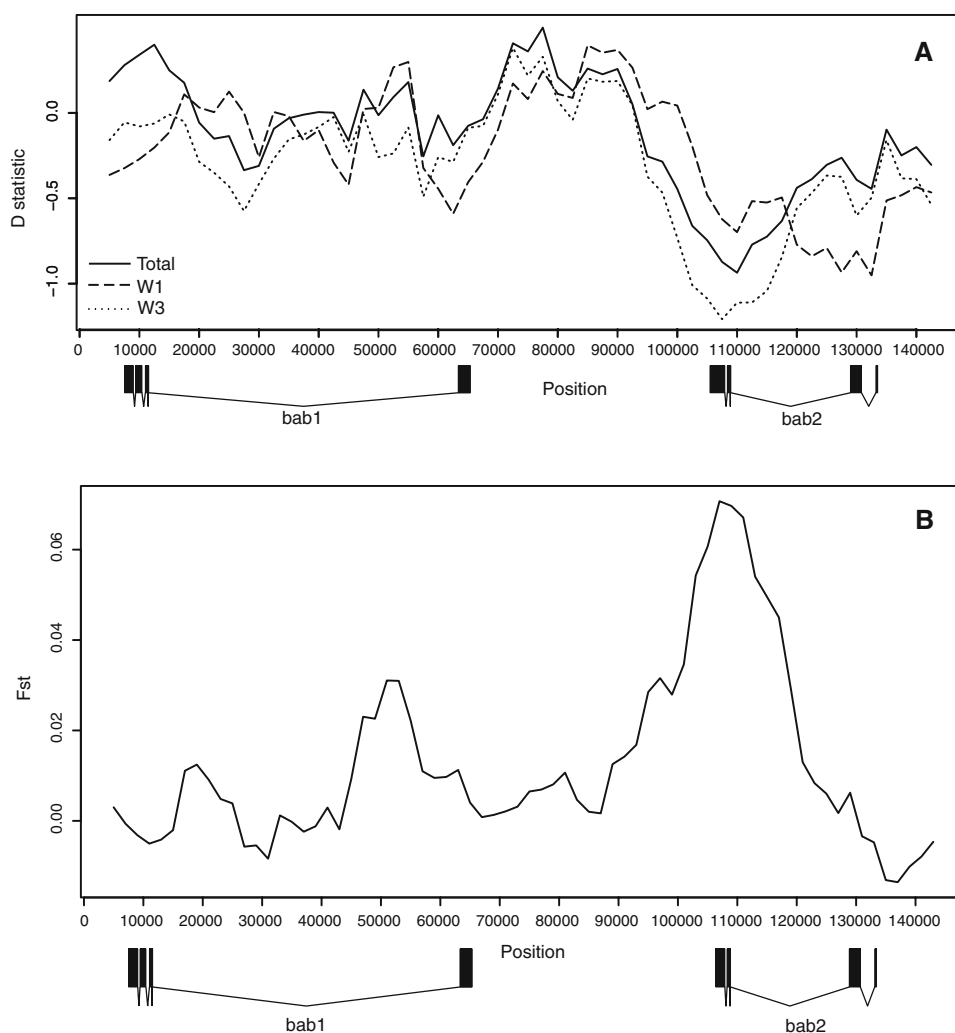
^b Number of polymorphisms identified

^c The number of polymorphisms that contain a single observation of the alternative allele

^d Waterson's Theta per 100 bp

^e The average pairwise diversity per 100 bp

Fig. 4 Comparison of the W1 and W3 collections. All analyses are in 10 kb windows offset by 2 kb. **a** Tajima's D statistic showing pooled collections (solid line), W1 alone (dashed line) and W3 alone (dotted line). **b** Population differentiation (F_{st}) between the W1 and W3 samples



differences suggest that selection acting on *bab2* may fluctuate over time. Further collections would be necessary to understand if these differences are the result of seasonal or yearly differences in selection for the abiotic environment or can be explained by some other process.

To investigate the role of selection in the evolution of *bab1* and *bab2* coding sequences, we used the McDonald–Kreitman (MK) test, which compares the ratio of synonymous (S) and non-synonymous (NS) nucleotide substitutions between and within species (McDonald and Kreitman 1991). To identify nucleotide substitutions that occurred specifically in the *D. melanogaster* lineage, we used the

genome sequences of *D. simulans* and *D. yakuba* to polarize the direction of change. We then compared the fixed substitutions that occurred on the *D. melanogaster* evolutionary lineage to polymorphisms segregating within the *D. melanogaster* population (Table 3). We found a significant skew in *bab1*, such that it has a lack of fixed NS changes and/or an excess of polymorphic NS substitutions (two tailed Fisher's exact test; $P < 0.05$). No such pattern is seen in the *bab2* gene or in the other predicted genes in the *bab* region (Table 3). The NS changes in *bab1* are distributed throughout the transcript, although none of these changes are found in the BTB (protein interaction) or

Table 3 McDonald–Kreitman test for genes in the *bab* region

Gene	Codons	Replacement (NS)		Silent (S)		P -value ^a
		Polymorphic	Fixed	Polymorphic	Fixed	
CG9205	199	0	0	10	7	1
<i>bab1</i>	791	15 (7)	1	56 (3)	28	0.034*
CG32334	135	5	1	6	0	1
<i>bab2</i>	828	13 (4)	6	79 (9)	30	0.784

^a P values are the results of a two-tailed Fisher's exact test. Values below $P = 0.05$ are indicated

BabCD (DNA binding) functional domains. This pattern could be due to functional constraint on the *bab1* coding region or balancing selection maintaining multiple alleles of *bab1*. To differentiate between these possibilities, we looked at how many of the polymorphic sites were represented by a single individual (singletons), which are more likely to be deleterious variants (Table 3). A large number of the NS polymorphisms in *bab1* are singletons (seven of 15) suggesting that *bab1* polymorphisms are under strong purifying selection preventing their fixation.

If the *bab* genes are functionally redundant, it is possible that deleterious alleles in one gene are compensated by a functional allele of the other gene. We tested for compensatory evolution between the *bab1* and *bab2* transcripts. We found no correlation between the number of low frequency (likely deleterious) alleles in the *bab1* and *bab2* coding regions ($P > 0.05$ for synonymous, non-synonymous, and total changes), nor do we find long-range LD between polymorphisms in the *bab1* and *bab2* transcripts. This suggests that there is not compensation between the *bab1* and *bab2* alleles.

Discussion

Duplicate genes persist in the genome due to the acquisition of new functions or the subdivision of the ancestral role. Over time, paralogous proteins may acquire subtle functional changes or gain entirely different biological activities (Hirth et al. 2001; Zhang et al. 2004). Alternatively, the proteins may share similar specificity while gene expression patterns diverge due to *cis*-regulatory changes, leading to the acquisition of different functional roles (Greer et al. 2000). The two mechanisms are not mutually exclusive, and both can operate on the same pair of paralogs. At the *bab* locus, the two duplicated genes have similar but non-identical expression patterns (Couderc et al. 2002) despite sharing at least some CREs (Williams et al. 2008). This suggests that some functions of these paralogs may experience shared constraints, while others may evolve independently.

Numerous studies have shown that duplicated genes diverge rapidly in expression (Gu et al. 2002; Makova and Li 2003; Gu et al. 2005), and that the rate of expression divergence is highest immediately after gene duplication and slows down over time (Jordan et al. 2004; Gu et al. 2005). This pattern is consistent with either directional selection (neo-functionalization model) or the relaxation of purifying selection (sub-functionalization model) acting during the early stages of gene divergence, and the relative contributions of these forces continue to be debated (Yu et al. 2003; Castillo-Davis et al. 2004; Jordan et al. 2004; Kondrashov and Kondrashov 2006). Generally, paralogous

genes are more likely to lose ancestral expression domains than to acquire new ones, indicating that sub-functionalization is probably more common than neo-functionalization (Oakley et al. 2006). Both models predict that, once duplicate genes acquire non-identical functions, they may come under different selective regimes.

In this study, we used a population genetic approach to assess the evolutionary forces acting on the *bab* paralogs. The patterns of sequence variation suggest that selective pressures vary across the *bab* locus. Two regions show indications of selection. First, a region near the 3' end of *bab2* (which includes *bab2* 3' exons, introns, and intergenic region) appears to experience directional selection (Fig. 3b, c). Furthermore, selection in this region may vary over time, as indicated by the difference between population samples collected in different years. Surprisingly, no CREs have been found in this region (Williams et al. 2008), although it remains possible that it contains regulatory elements that modulate transcriptional activity but cannot function independently in transgenic assays. Future analysis is required to determine whether the coding or non-coding DNA is driving this signature of selection. The second region that appears to be under selection is located in the large intron of *bab1* (Fig. 3c). This region contains the CRE that controls female specific expression of *bab* in the abdominal epidermis (Williams et al. 2008), suggesting that selection may be acting on the sexually dimorphic pigmentation of *D. melanogaster*. Furthermore, a recent study found that this same region was differentiated between northern and southern *D. melanogaster* populations in North America and Australia (Turner et al. 2008). In the coding regions, *bab1* exhibits stronger selective constraint than *bab2*. One possible explanation is that the two proteins have somewhat different functional activities despite being expressed in largely overlapping patterns.

Given our data, it seems that the *bab* homologs are most likely maintained due to sub-functionalization. Previous work on the *bab* locus has shown that both *bab* genes are expressed in the same tissues during development (Couderc et al. 2002). This suggests that both genes probably maintain similar functions as the ancestral *bab* gene. We have found that the coding and non-coding DNA show differences in sequence evolution. Thus, within the ancestral functions it is likely that the *bab* genes have divided their roles such that both are indispensable and thus maintained.

Several recent studies have used comparative genomic approaches to examine the role of selection in the evolution of duplicate genes. Such analyses are based on variation in the rate of expression divergence over time (Jordan et al. 2004; Gu et al. 2005), across phylogenetic lineages (Shiu et al. 2006), or on the correlation between the rate of expression and sequence divergence (Yu et al. 2003;

Castillo-Davis et al. 2004). However, these long-term evolutionary patterns are consistent with either selective or neutral explanations (Castillo-Davis et al. 2004; Jordan et al. 2004; Kondrashov and Kondrashov 2006), and are best suited for detecting selection at the genome-wide level rather than individual loci. A population-genetic approach brings an alternative perspective to this question, since it is explicitly designed to test for selection acting on specific DNA sequences. As genome-wide analyses of intraspecific variation become possible (Begun et al. 2007), an integration of population-genetic and comparative-genomic approaches will shed new light on the relative importance of positive selection and neutral changes in the maintenance and evolution of paralogous genes.

Acknowledgements We would like to thank Anna Ustaszewska and Danielle Tufts for technical help, and Tina Hu and Jennifer Brisson for comments on the manuscript. This work was supported by NSF grant DEB-0548991 to AK and SN.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Adams MDSE, Celniker RA, Holt CA, Evans JD, Gocayne PG, Amanatides SE, Scherer PW, Li RA, Hoskins RF, Galle RA, George SE, Lewis S, Richards M, Ashburner SN, Henderson GG, Sutton JR, Wortman MD, Yandell Q, Zhang LX, Chen RC, Brandon YH, Rogers RG, Blazej M, Champe BD, Pfeiffer KH, Wan C, Doyle EG, Baxter G, Helt CR, Nelson GL, Gabor JF, Abril A, Agbayani HJ, An C, Andrews-Pfannkoch D, Baldwin RM, Ballew A, Basu J, Baxendale L, Bayraktaroglu EM, Beasley KY, Beeson PV, Benos BP, Berman D, Bhandari S, Bolshakov D, Borkova MR, Botchan J, Bouck P, Brokstein P, Brottier KC, Burtis DA, Busam H, Butler E, Cadieu A, Center I, Chandra JM, Cherry S, Cawley C, Dahlke LB, Davenport P, Davies B, de Pablos A, Delcher Z, Deng AD, Mays I, Dew SM, Dietz K, Dodson LE, Doup M, Downes S, Dugan-Rocha BC, Dunkov P, Dunn KJ, Durbin CC, Evangelista C, Ferraz S, Ferriera W, Fleischmann C, Fosler AE, Gabrielian NS, Garg WM, Gelbart K, Glasser A, Glodek F, Gong JH, Gorrell Z, Gu P, Guan M, Harris NG, Harris D, Harvey TJ, Heiman JR, Hernandez J, Houck D, Hostin KA, Houston TJ, Howland MH, Wei C, Ibegwam M, Jalali F, Kalush GH, Karpen Z, Ke JA, Kennison KA, Ketchum BE, Kimmel CD, Kodira C, Kraft S, Kravitz D, Kulp Z, Lai P, Lasko Y, Lei AA, Levitsky J, Li Z, Li Y, Liang X, Lin X, Liu B, Matthei TC, McIntosh MP, McLeod D, McPherson G, Merkulov NV, Milshina C, Mobarry J, Morris A, Moshrefi SM, Mount M, Moy B, Murphy L, Murphy DM, Muzny DL, Nelson DR, Nelson KA, Nelson K, Nixon DR, Nusskern JM, Pacleb M, Palazzolo GS, Pittman S, Pan J, Pollard V, Puri MG, Reese K, Reinert K, Remington RD, Saunders F, Scheeler H, Shen BC, Shue I, Siden-Kiamos M, Simpson MP, Skupski T, Smith E, Spier AC, Spradling M, Stapleton R, Strong E, Sun R, Svirskas C, Tector R, Turner E, Venter AH, Wang X, Wang ZY, Wang DA, Wassarman GM, Weinstock J, Weissenbach SM, WilliamsWoodage TKC, Worley D, Wu S, Yang QA, Yao J, Ye RF, Yeh JS, Zaveri M, Zhan G, Zhang Q, Zhao L, Zheng XH, Zheng FN, Zhong W, Zhong X, Zhou S, Zhu X, Zhu HO, Smith RA, Gibbs EW, Myers G, Rubin M, Venter JC (2000) The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195
- Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL, Westerfield M, Ekker M, Postlethwait JH (1998) Zebrafish hox clusters and vertebrate genome evolution. Science 282:1711–1714
- Amores A, Suzuki T, Yan YL, Pomeroy J, Singer A, Amemiya C, Postlethwait JH (2004) Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. Genome Res 14:1–10
- Andolfatto P (2005) Adaptive evolution of non-coding DNA in *Drosophila*. Nature 437:1149–1152
- Bardot O, Godt D, Laski FA, Couderc JL (2002) Expressing UAS-bab1 and UAS-bab2: a comparative study of gain-of-function effects and the potential to rescue the bric a brac mutant phenotype. Genesis 34:66–70
- Barmina O, Kopp A (2007) Sex-specific expression of a HOX gene associated with rapid morphological evolution. Dev Biol 311:277–286
- Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, Pachter L, Myers E, Langley CH (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol 5:e310
- Brisson JA, De Toni DC, Duncan I, Templeton AR (2005) Abdominal pigmentation variation in *Drosophila polymorpha*: geographic variation in the trait, and underlying phylogeography. Evolution Int J org Evolution 59:1046–1059
- Castillo-Davis CI, Hartl DL, Achaz G (2004) cis-Regulatory and protein evolution in orthologous and duplicate genes. Genome Res 14:1530–1536
- Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–1303
- Chen W, Zollman S, Couderc JL, Laski FA (1995) The BTB domain of bric a brac mediates dimerization in vitro. Mol Cell Biol 15:3424–3429
- Chu J, Dong PD, Panganiban G (2002) Limb type-specific regulation of bric a brac contributes to morphological diversity. Development 129:695–704
- Couderc JL, Godt D, Zollman S, Chen J, Li M, Tiong S, Cramton SE, Sahut-Barnola I, Laski FA (2002) The bric a brac locus consists of two paralogous genes encoding BTB/POZ domain proteins and acts as a homeotic and morphogenetic regulator of imaginal development in *Drosophila*. Development 129:2419–2433
- Cresko WA, Yan YL, Baltrus DA, Amores A, Singer A, Rodriguez-Mari A, Postlethwait JH (2003) Genome duplication, subfunction partitioning, and lineage divergence: Sox9 in stickleback and zebrafish. Dev Dyn 228:480–489
- Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186–194
- Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175–185
- Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413
- Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545
- Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147:915–925
- Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133:693–709

- Gibert P, Moreteau B, Moreteau JC, David JR (1996) Growth temperature and adult pigmentation in two *Drosophila* sibling species: an adaptive convergence of reaction norms in sympatric populations? *Evolution Int J org Evolution* 50:2346–2353
- Godt D, Laski FA (1995) Mechanisms of cell rearrangement and cell recruitment in *Drosophila* ovary morphogenesis and the requirement of bric a brac. *Development* 121:173–187
- Godt D, Couderc JL, Cramton SE, Laski FA (1993) Pattern formation in the limbs of *Drosophila*: bric a brac is expressed in both a gradient and a wave-like pattern and is required for specification and proper segmentation of the tarsus. *Development* 119:799–812
- Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Res* 8:195–202
- Greer JM, Puetz J, Thomas KR, Capecchi MR (2000) Maintenance of functional equivalence during paralogous Hox gene evolution. *Nature* 403:661–665
- Gu Z, Nicolae D, Lu HH, Li WH (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet* 18:609–613
- Gu X, Zhang Z, Huang W (2005) Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc Natl Acad Sci USA* 102:707–712
- Hirth F, Loop T, Egger B, Miller DF, Kaufman TC, Reichert H (2001) Functional equivalence of Hox gene products in the specification of the tritocerebrum during embryonic brain development of *Drosophila*. *Development* 128:4781–4788
- Holland PWH (1999) Gene duplication: past, present and future. *Semin Cell Dev Biol* 10:541–547
- Holland PWH, Garcia-Fernandez J, Williams NA, Sidow A (1994) Gene duplications and the origins of vertebrate development. *Dev Suppl* 125–133
- Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from DNA-sequence data. *Genetics* 132:583–589
- Jordan IK, Wolf YI, Koonin EV (2004) Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol Biol* 4:22
- Kondrashov FA, Kondrashov AS (2006) Role of selection in fixation of gene duplications. *J Theor Biol* 239:141–151
- Kopp A, Duncan I, Carroll SB (2000) Genetic control and evolution of sexually dimorphic characters in *Drosophila*. *Nature* 408:553–559
- Kopp A, Graze RM, Xu SZ, Carroll SB, Nuzhdin SV (2003) Quantitative trait loci responsible for variation in sexually dimorphic traits in *Drosophila melanogaster*. *Genetics* 163:771–787
- Lehmann M, Siegmund T, Lintermann KG, Korge G (1998) The pipsqueak protein of *Drosophila melanogaster* binds to GAGA sequences through a novel DNA-binding domain. *J Biol Chem* 273:28504–28509
- Lewontin RC (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49:49–67
- Lours C, Bardot O, Godt D, Laski FA, Couderc JL (2003) The *Drosophila melanogaster* BTB proteins bric a brac bind DNA through a composite DNA binding domain containing a pipsqueak and an AT-Hook motif. *Nucleic Acids Res* 31:5389–5398
- Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154:459–473
- Makova KD, Li WH (2003) Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res* 13:1638–1645
- McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351:652–654
- Ng CS, Kopp A (2008) Sex combs are important for male mating success in *Drosophila melanogaster*. *Behav Genet* 38:195–201
- Oakley TH, Ostman B, Wilson AC (2006) Repression and loss of gene expression outpaces activation and gain in recently duplicated fly genes. *Proc Natl Acad Sci USA* 103:11637–11641
- Ohno S (1970) *Evolution by gene duplication*. Springer-Verlag, Heidelberg, Germany
- Prince VE, Pickett FB (2002) Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet* 3:827–837
- Randsholt NB, Santamaria P (2008) How *Drosophila* change their combs: the Hox gene sex combs reduced and sex comb variation among *Sophophora* species. *Evol Dev* 10:121–133
- Reeves R, Nissen MS (1990) The A.T-DNA-binding domain of mammalian high mobility group I chromosomal proteins. A novel peptide motif for recognizing DNA structure. *J Biol Chem* 265:8573–8582
- Sahut-Barnola I, Godt D, Laski FA, Couderc JL (1995) *Drosophila* ovary morphogenesis: analysis of terminal filament formation and identification of a gene required for this process. *Dev Biol* 170:127–135
- Shiu SH, Byrnes JK, Pan R, Zhang P, Li WH (2006) Role of positive selection in the retention of duplicate genes in mammalian genomes. *Proc Natl Acad Sci USA* 103:2232–2236
- Smit A, Hubley R, Green P (1996–2004) RepeatMasker Open-3.0
- Stephens M, Sloan JS, Robertson PD, Scheet P, Nickerson DA (2006) Automating sequence-based detection and genotyping of SNPs from diploid samples. *Nat Genet* 38:375–381
- Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595
- Thornton K (2003) Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* 19:2325–2327
- Turner TL, Levine MT, Eckert ML, Begun DJ (2008) Genomic analysis of adaptive differentiation in *Drosophila melanogaster*. *Genetics* 179:455–473
- Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* 7:256–275
- Williams TM, Selegue JE, Werner T, Gompel N, Kopp A, Carroll SB (2008) The regulation and evolution of a genetic switch controlling sexually dimorphic traits in *Drosophila*. *Cell* 134:610–623
- Yu WP, Brenner S, Venkatesh B (2003) Duplication, degeneration and subfunctionalization of the nested synapsin-Timp genes in *Fugu*. *Trends Genet* 19:180–183
- Zeng K, Fu YX, Shi S, Wu CI (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174:1431–1439
- Zhang J, Dean AM, Brunet F, Long M (2004) Evolving protein functional diversity in new genes of *Drosophila*. *Proc Natl Acad Sci USA* 101:16246–16250
- Zollman S, Godt D, Prive GG, Couderc JL, Laski FA (1994) The BTB domain, found primarily in zinc finger proteins, defines an evolutionarily conserved family that includes several developmentally regulated genes in *Drosophila*. *Proc Natl Acad Sci USA* 91:10717–10721